PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-185196
(43)Date of publication of application : 16.07.1996

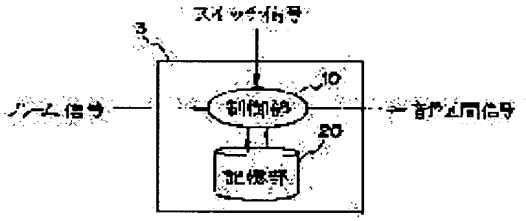
(51)Int.Cl. G10L 3/00
G10L 3/00

(21)Application number : 06-329161
(22)Date of filing : 28.12.1994

(71)Applicant : SONY CORP
(72)Inventor : MINAMINO KATSUKI
ISHII KAZUO
OGAWA HIROAKI

(54) DEVICE FOR DETECTING SPEECH SECTION

(57)Abstract:
PURPOSE: To detect a speech section without lacking a word head even if utterance is performed with timing at which a switch is not yet pressed by extracting only one speech section from a stored input speech signal.
CONSTITUTION: A speech section detecting device 3 has a control section 10 and a storage section 20 such as a memory. And the control section 10 always discriminates whether it is voice or voiceless for a speech signal made a digital signal by an A/D converter from a microphone, on the other hand, only one speech section is extracted from a speech signal stored in the memory part 20 in a wider range than a detection range specified by switch operation of a section specifying switch and outputted. Also, the memory part 20 always preserves an input speech signal of some fixed time. Therefore, a speech section can be detected without lacking a word head even if utterance is performed with timing at which a switch is not yet pressed.



LEGAL STATUS

[Date of request for examination]	11.01.2001
[Date of sending the examiner's decision of rejection]	30.06.2003
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	
[Date of registration]	
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平8－185196

(43)公開日　平成８年(1996)７月16日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ１０Ｌ　3/00	５１３　Ａ			
	Ｚ			
	５７１　Ｇ			

審査請求　未請求　請求項の数15　ＯＬ　（全 13 頁）

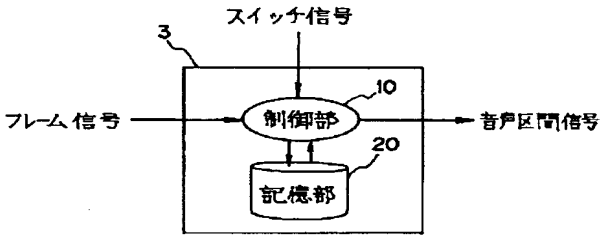
(21)出願番号　特願平6－329161

(22)出願日　平成６年(1994)12月28日

(71)出願人　000002185
ソニー株式会社
東京都品川区北品川６丁目７番35号
(72)発明者　南野　活樹
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(72)発明者　石井　和夫
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(72)発明者　小川　浩明
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(74)代理人　弁理士　小池　晃　（外２名）

(54)【発明の名称】　音声区間検出装置

(57)【要約】
【構成】　制御部１０は、フレーム音声信号に対して、有音声と無音声とを常に判定しながら、一方でスイッチ操作によるスイッチ信号で指定される検出範囲よりも広い範囲で記憶部２０に記憶されている上記音声信号から一つだけ音声区間を抽出し、出力する。
【効果】　スイッチが押されるより早いタイミングで発声が行われた場合でも、語頭を欠くことなく音声区間を検出できる。



1

【特許請求の範囲】

【請求項１】　音声信号が入力され、検出範囲を指定するスイッチ操作に応じて音声区間を抽出して出力する音声区間検出装置において、

上記入力音声信号を記憶する記憶手段と、

上記スイッチ操作で指定される検出範囲よりも広い範囲で上記記憶手段に記憶されている上記入力音声信号から一つだけ音声区間を抽出し、出力する制御手段とを有することを特徴とする音声区間検出装置。

【請求項２】　上記制御手段は、上記記憶手段から上記音声区間を抽出し、該音声区間を入力信号から一定時間遅らせて送信することを特徴とする請求項１記載の音声区間検出装置。

【請求項３】　上記制御手段は、上記入力された音声信号を周波数分析して信号パワースペクトルを求め、環境雑音のパワースペクトルの定数倍に対する該信号パワースペクトルの大小に応じて入力時の雑音を除去することを特徴とする請求項１記載の音声区間検出装置。

【請求項４】　上記制御手段は、上記信号パワースペクトルの積算値が所定のしきい値以上ならば有音声、未満ならば無音声と判定することを特徴とする請求項３記載の音声区間検出装置。

【請求項５】　上記制御手段は、上記有音声と判定される区間が所定長以上続いたときに、これを音声区間として検出し、その後に上記無音声と判定される区間が所定長以上続いたときに、上記音声区間が終了したと見なすことを特徴とする請求項４記載の音声区間検出装置。

【請求項６】　上記制御手段は、短い有音声部の後に無音声部が続き、その後再び有音声部が続くような場合、上記短い有音声部と上記無音声部の長さの比に応じて上記短い有音声部を音声区間とするか否かを判定することを特徴とする請求項５記載の音声区間検出装置。

【請求項７】　上記制御手段は、一定の長さの有音声部の後に、無音声部が続き、その後短い有音声が続くような場合、上記無音声部と上記短い有音声部の長さの比に応じて上記短い有音声部を音声区間とするか否かを判定することを特徴とする請求項５記載の音声区間検出装置。

【請求項８】　上記制御手段は、上記音声区間と判定された区間の前後にマージンを付加して、音声区間を引き延ばすことを特徴とする請求項５、６又は７記載の音声区間検出装置。

【請求項９】　上記制御手段は、上記有音声／無音声の判定のためのしきい値や、上記マージンのようなパラメータを環境雑音の平均エネルギーに応じて変動させることを特徴とする請求項４又は８記載の音声区間検出装置。

【請求項１０】　上記制御手段は、上記パラメータと上記環境雑音の平均エネルギーとの関係を比例関係として予め決めておき、さらにそのパラメータの上限と下限も

2

決めておくことで、上記環境雑音の平均エネルギーから上記パラメータを決定することを特徴とする請求項９記載の音声区間検出装置。

【請求項１１】　上記制御部は、環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを無音声区間で更新することを特徴とする請求項９又は１０記載の音声区間検出装置。

【請求項１２】　上記制御部は、連続して無音声と判定され続け、かつ更新前の環境雑音の平均エネルギーから緩やかに変化するようなエネルギーを持つ区間においてのみ、環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを更新することを特徴とする請求項１１記載の音声区間検出装置。

【請求項１３】　上記制御部は、一定時間以上有音声と判定され続けるような場合には、強制的に環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを更新することを特徴とする請求項９、１０、１１又は１２記載の音声区間検出装置。

【請求項１４】　上記制御部は、上記音声区間の抽出を常に行うことを特徴とする請求項１記載の音声区間検出装置。

【請求項１５】　上記制御部は、上記環境雑音の平均エネルギー、上記環境雑音の平均パワースペクトル及び上記パラメータの更新を常に行うことを特徴とする請求項９、１０、１１、１２又は１３記載の音声区間検出装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、入力された音声信号から、有音声部と無音声部を識別し、スイッチ操作と連動させて必要な音声区間のみを検出する音声区間検出装置に関する。

【０００２】

【従来の技術】従来、音声区間を検出するには、音声信号のパワーがあるしきい値を越えれば有音声部、そのしきい値を越えなければ無音声部と判定し、ある一定長以上の有音声部が得られたときに、これを音声区間として検出するという方法がある。また、有音声部と無音声部の判別に、零交差数を用いたり、周波数分解したスペクトルパワーを用いたりする方法、さらに、これらを組み合わせて判別する方法もある。また、環境雑音の平均エネルギーに応じてしきい値を変動させたり、環境雑音の平均スペクトルと入力信号のスペクトルとの類似性や相関などを調べることで、雑音環境下でも音声区間が頑強に検出できるようにした方法等もある。これらの方法をそのまま音声認識システムや音声通信システムの前処理として適用した場合、うまく音声区間を検出することができれば、後段の処理が正常に動作することが期待できるわけだが、実際には、その音声区間の前後に発生される不要な音声も検出されてしまい、システム全体の動作

に悪影響を及ぼすという問題が生じる。例えば、地名を認識する音声認識システムでは、話者が「しながわ」と発声すれば、その発声に対して認識を行うようにしたいわけだが、その発声の前に「えーっと、」などと発声した場合にはこれも音声区間として検出されてしまい、その発声に対して認識が行われることになってしまう。同様に、「しながわ」と発声した後しばらくしてから「・・・だよね？」などと横にいる話し手に話しかけた場合、話者はシステムに話したつもりはなくても、やはりこれが音声区間として検出されてしまい、その発声に対しても認識が行われることになってしまう。

【０００３】これらの問題は、音声認識システムがユーザの発声した音声の中で、必要な音声区間だけを選ぶことができないことに起因する。

【０００４】そこで、スイッチを付け、必要な音声区間をユーザが直接指定するようにすることで、この問題を解決する方法が考えられる。一つは、スイッチが押されている間を音声区間として抽出する方法、一つは、スイッチが一度押された後、一定時間の間に発声された音声区間を前述の有音声と無音声の判別方法を用いて抽出する方法などである。

【０００５】

【発明が解決しようとする課題】ところで、これらの方法においては、ユーザが発声より少しでも遅れてスイッチを押した場合に、抽出すべき音声区間の最初の部分が欠けてしまうという問題が生じてしまう。また、環境雑音の平均エネルギーに応じてしきい値を変動させたり、環境雑音の平均スペクトルを用いるなどするような雑音環境下にも適応できる区間検出方法を用いた場合、従来は、その環境雑音を抽出するためにスイッチを押してから一定時間環境雑音を観測するということを行うため、その間ユーザは話せないなどの煩わしさを伴うという問題があった。

【０００６】本発明は、上記実情に鑑みてなされたものであり、ユーザにスイッチを押すタイミングと発声のタイミングを意識させるような煩わしさを不要とさせ、必要な音声区間だけをうまく検出し、かつ雑音環境下でも安定した音声区間検出を実現できる音声区間検出装置の提供を目的とする。

【０００７】

【課題を解決するための手段】本発明に係る音声区間検出装置は、音声信号が入力され、検出範囲を指定するスイッチ操作に応じて音声区間を抽出して出力する音声区間検出装置において、上記入力音声信号を一定時間分記憶する記憶手段と、上記スイッチ操作で指定される検出範囲よりも広い範囲で上記記憶手段に記憶されている上記入力音声信号から一つだけ音声区間を抽出し、出力する制御手段とを有することにより上記課題を解決する。

【０００８】この場合、上記制御手段は、上記記憶手段から上記音声区間を抽出し、該音声区間を入力信号から一定時間遅らせて送信する。

【０００９】また、上記制御手段は、上記入力された音声信号を周波数分析して信号パワースペクトルを求め、環境雑音のパワースペクトルの定数倍に対する該信号パワースペクトルの大小に応じて入力時の雑音を除去する。

【００１０】また、上記制御手段は、上記信号パワースペクトルの積算値が所定のしきい値以上ならば有音声、未満ならば無音声と判定する。

【００１１】また、上記制御手段は、上記有音声と判定される区間が所定長以上続いたときに、これを音声区間として検出し、その後に上記無音声と判定される区間が所定長以上続いたときに、上記音声区間が終了したと見なす。

【００１２】また、上記制御手段は、短い有音声部の後に無音声部が続き、その後再び有音声部が続くような場合、上記短い有音声部と上記無音声部の長さの比に応じて上記短い有音声部を音声区間とするか否かを判定する。

【００１３】また、上記制御手段は、一定の長さの有音声部の後に、無音声部が続き、その後短い有音声が続くような場合、上記無音声部と上記短い有音声部の長さの比に応じて上記短い有音声部を音声区間とするか否かを判定する。

【００１４】また、上記制御手段は、上記音声区間と判定された区間の前後にマージンを付加して、音声区間を引き延ばす。

【００１５】また、上記制御手段は、上記有音声／無音声の判定のためのしきい値や、上記マージンのようなパラメータを環境雑音の平均エネルギーに応じて変動させる。

【００１６】また、上記制御手段は、上記パラメータと上記環境雑音の平均エネルギーとの関係を比例関係として予め決めておき、さらにそのパラメータの上限と下限も決めておくことで、上記環境雑音の平均エネルギーから上記パラメータを決定する。

【００１７】また、上記制御部は、環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを無音声区間で更新する。

【００１８】また、上記制御部は、連続して無音声と判定され続け、かつ更新前の環境雑音の平均エネルギーから緩やかに変化するようなエネルギーを持つ区間においてのみ、環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを更新する。

【００１９】また、上記制御部は、一定時間以上有音声と判定され続けるような場合には、強制的に環境雑音の平均エネルギーや環境雑音の平均パワースペクトルを更新する。

【００２０】また、上記制御部は、上記音声区間の抽出を常に行う。

5

【００２１】また、上記制御部は、上記環境雑音の平均エネルギー、上記環境雑音の平均パワースペクトル及び上記パラメータの更新を常に行う。

【００２２】

【作用】有音声と無音声の判別による音声区間判定を常に行いながら、その判定結果と一定時間前までの入力信号を記憶しておくことで、スイッチが押されるより早いタイミングで発声が行われた場合でも、語頭を欠くことなく音声区間を検出することが可能と成る。又、一つのスイッチ指定区間に対して一つの音声区間だけを検出し送信するようにしたことで、話者が発声の少し前や、少し後に関係のない言葉を発声したとしても、これを音声区間として検出することがなくなるため、例えば音声認識装置に適用した場合などに誤動作を起こしにくくなる。

【００２３】さらに、入力信号を周波数分析して得られるパワースペクトルを用いること、及び、環境雑音の平均エネルギーに適応させてパラメータを更新することなどを行うことで、環境雑音下においても安定した音声区間検出が行えるようにすると共に、その適応化を音声区間の判定と平行して常に行っておくようにすることで、スイッチを押してから雑音環境の観測を行う必要が特になくなるため、ユーザが発声するのを待たされることもなくなる。つまり、ユーザは、単純にスイッチを押せば一回だけ発話することができるということだけを覚えておけばよく、その発話とスイッチのタイミングをそれほど意識しなくてよいものとなる。しかも、システムは、必要な音声区間だけを検出することが可能となる。

【００２４】そして、この音声区間検出装置を音声認識システムや音声通信システムに組み込めば、入力信号のなかから認識処理を行うべき音声区間や送信すべき音声区間を抽出することができるようになる。

【００２５】

【実施例】以下、本発明に係る音声区間検出装置の実施例について説明する。この実施例となる音声区間検出装置は、話者にスイッチを押すタイミングと発声のタイミングとをあまり意識させることなく、必要な音声区間を検出できる。この音声区間検出装置は、例えば図１に示すような音声認識システムに適用され、検出した音声区間は音声認識のために使われる。先ず、本実施例の音声区間検出装置を説明する前に、図１に示す音声認識システムについて説明する。

【００２６】マイクロホン１で収音された音声信号は、Ａ／Ｄ変換器２でディジタル信号に変換され、本実施例の音声区間検出装置３に供給される。この音声区間検出装置３には、話者が押した区間指定スイッチ４からのオン、オフ信号も供給される。そして、音声区間検出装置３は、上記ディジタル信号と、上記オン、オフ信号を用いることで、必要な音声区間の音声信号を検出し、これを音声認識部５に送る。音声認識部５は、音声区間検出

6

装置３が検出した音声信号に対して認識処理、すなわち音響分析やベクトル量子化などによる特徴量の抽出と、ダイナミックプログラミング（以下、ＤＰという。）マッチングやＨＭＭなどによるスコア計算を行ない、その認識結果を出力する。このような音声認識システムは、さまざまな機器のコントロールを音声で行う場合や、キーボードなどに変わる入力手段の一つとして、広く用いられることが期待できる。

【００２７】このような音声認識システムにおいて、話者にあまり煩わしさを与えることなく、必要な音声区間を抽出することが重要となる。このため、本実施例となる音声区間検出装置３が必要となる。

【００２８】この音声区間検出装置３は、図２に示すように、制御部１０と、例えばメモリのような記憶部２０とを有して成る。制御部１０は、上記図１に示したＡ／Ｄ変換器２でディジタル信号とされたマイクロホン１からの音声信号に対して、有音声と無音声とを常に判定しながら、一方で区間指定スイッチ４のスイッチ操作で指定される検出範囲よりも広い範囲で記憶部２０に記憶されている上記音声信号から一つだけ音声区間を抽出し、出力する。また、制御部１０は、環境雑音の観測、内部パラメータの更新なども行うようにする。記憶部２０は、ある一定時間分の入力音声信号を常に保存しておく。

【００２９】制御部１０に入力されるＡ／Ｄ変換器２からのディジタル入力音声信号は、サンプリングされたデータであり、例えば１２８サンプルをまとめて１フレームとされ、１フレームずつ順に送られて来るものとする。なお、以下では、時刻ｔに制御部１０に入力されるフレーム信号をF_tと表す。

【００３０】この制御部１０は、図３に示すフローチャートに基づいた動作を行う。

【００３１】先ず、このフローチャートが開始されると、この制御部１０は、ステップＳ１に示すように、上記フレーム信号F_tを記憶部２０に保存する。ここで、この記憶部２０には、常にｋフレーム前までの入力信号を記憶しておくようにする。すなわち、記憶部２０は、時刻ｔにはｋ時刻前までのフレーム信号F_t、F_{t-1}、F_{t-2}・・・F_{t-k}を記憶している。

【００３２】次に、この制御部１０は、ステップＳ２からステップＳ５までに示すような動作を行い、時刻ｔの入力フレーム信号F_tに音声が含まれるかどうか、つまり有音声か無音声かを判別する。ステップＳ２からステップＳ５までに示す動作は、周波数分析を用いる方法を適用している。

【００３３】ステップＳ２では、入力されたフレーム信号F_tの周波数分析を行う。具体的には、入力されたサンプル信号に、ハミングウィンドウをかけ、高速フーリエ変換（以下、ＦＦＴという。）を施すことで、パワースペクトルを求めたり、バンドパスフィルタを用いて各

7

帯域毎のパワースペクトルを抽出する。

【００３４】ステップＳ３では、環境雑音の平均パワースペクトルを用いて、雑音除去を行う。具体的に、雑音除去後のパワースペクトルＸ（ω）は、入力信号のパワースペクトルをＳ（ω）、環境雑音の推定パワースペクトルをＮ（ω）、オフセットをＲ、環境雑音の推定パワースペクトルＮ（ω）にかける重み係数をαとすると、

【００３５】

【数１】

$$X(\omega)=\begin{cases}S(\omega)+R & S(\omega)\geq\alpha\cdot N(\omega)\\ R & S(\omega)<\alpha\cdot N(\omega)\end{cases}\quad\cdots(1)$$

【００３６】のように求められる。したがって、この制御部１０は、ある帯域ωにおいて、入力信号のパワースペクトルＳ（ω）が環境雑音の推定パワースペクトルＮ（ω）のα倍以上ならば、その帯域ωには雑音以外のパワースペクトルが多く含まれていると見なしオフセットＲを付加して上記入力信号のパワースペクトルＳ（ω）をそのまま残し、逆に、もし上記入力信号のパワースペクトルＳ（ω）が上記環境雑音の推定パワースペクトルＮ（ω）のα倍よりも小さいならば、その帯域には雑音しか含まれていないと見なしオフセットＲのみとすることで、雑音のパワースペクトルを除去している。これにより、環境雑音と周波数成分の異なる信号がある程度の大きさで入力されれば、その周波数成分の存在する帯域において、入力信号のパワースペクトルＳ（ω）が残ることになる。例えば、静かな環境、ファンノイズの環境、自動車走行雑音の環境など、さまざまな雑音環境下において音声が発声された場合、環境雑音が多少大きなときでも、音声帯域と環境雑音の帯域の相違から、雑音除去後のパワースペクトルＸ（ω）の音声帯域における入力信号のパワースペクトルＳ（ω）は除去されずに残ることになる。

【００３７】そこで、制御部１０は、ステップＳ４に示すように、ステップＳ３で得られたパワースペクトルＸ（ω）からエネルギーＥを、

【００３８】

【数２】

$$E=10\log_{10}\left(\sum_{\omega}X(\omega)\right)\quad\cdots(2)$$

【００３９】のように求めた後、このエネルギーＥがあるしきい値ｒより大きいか否かを判定し、ステップＳ５に示すように有音声／無音声の判定を行っている。上記（２）式のエネルギーＥが上記しきい値ｒ以上ならば有音声、上記しきい値未満ならば無音声と判定する。ここで、しきい値ｒとしては、例えば、

【００４０】

【数３】

8

$$r=10\log_{10}\left(\sum_{\omega}R\right)+\delta\quad\cdots(3)$$

【００４１】のような一定値を用いたり、あるいは後述するような、環境雑音の平均エネルギーに適応して変動させた値を用いればよい。ただし、δは定数である。

【００４２】なお、ここでは、有音声／無音声判定部の判定方法として、周波数分析により得たパワースペクトルを用いる方法を述べたが、これに代わる方法として、従来のパワーや零交差数などによる判定方法を用いてもよい。

【００４３】次に、制御部１０は、ステップＳ６に示すように、音声区間の判定を行う。音声区間の判定は、上述したような有音声と無音声の判定に基づき、有音声部が最低ｍフレーム（例えば、３０フレーム）以上続く場合の有音声部の始端から終端までを、基本的に音声区間として判定するようにする。また、ｎフレーム（例えば２５フレーム）以上の無音声部が続けば、これを無音声区間と判定し、上記音声区間が終了したと見なす。

【００４４】この時、音声区間の始端をできるだけ早く検出するため、図４の（Ａ）に示すように、フレームＦｔが入力されたときに、フレームＦｔ-ｍが音声区間のフレームであるかどうかを判定し、その結果を記憶部２０に書き込むようにする。すなわち、時刻ｔに右端のフレームＦｔが入力されたときに、ｍ時刻前のフレームＦｔ-ｍが音声区間かどうかを判定し、その結果を記憶部２０に書き込ませる。図４の（Ａ）の斜線部分は、有音声部を表す。また、後述する図４の（Ｂ）乃至図４の（Ｇ）において、示される斜線部分も有音声部を表し、それを除く部分は無音声部を表す。

【００４５】以下、図４の（Ｂ）乃至図４の（Ｇ）を参照しながらステップＳ６の音声区間判定の動作の詳細を説明する。図４の（Ｂ）は、音声が入力されておらず、無音声区間中にさらに無音声フレームが入力された場合であり、Ｆｔ-ｍは無音声区間として判定される。図４の（Ｃ）は、音声が入力され始め、有音声フレームが数フレーム連続して入力された場合であるが、まだｍフレーム以上の有音声部が続いていないので、やはりＦｔ-ｍは無音声区間として判定される。図４の（Ｄ）は、フレームＦｔが有音声部として判定され、有音声部がちょうどｍフレーム続いたときを示し、このときフレームＦｔ-ｍが音声区間の始まりとして検出される。その後、図４の（Ｅ）のように、しばらく入力フレームが有音声と判定された場合、フレームＦｔ-ｍは音声区間として判定され続ける。しばらくして、音声の入力が終了すれば、図４の（Ｆ）のように無音声フレームが入力され始める。しかし、フレームＦｔ-ｍはまだ音声区間として判定されたままとなる。そして、図４の（Ｇ）のようにｎフレームの無音声フレームが続けて入力された場合に、音声区間の終了が検出される。ただし、図４の（Ｇ）はｎ≦ｍの

9

場合を示しており、この場合は、その後m－nフレームが音声区間として判定されることになる。そしてしばらくして、再び図４の（B）に示す状態に戻る。

【００４６】また、n＞mの場合は、図４の（G）が少し変わり、音声の入力が終了し、連続してmフレーム無音声フレームが入力されても、フレームF_{t-m}は音声区間として判定されたままとなる。そこで、nフレームの連続した無音声フレームが入力された時点で、F_{t-m}を無音声区間と判定し、同時に、そのn－mフレーム前までの音声区間と判定されたフレームF_{t-m-1}、F_{t-m-2}・・・F_{t-n}の判定結果を無音声区間と書き換えるようにする。

【００４７】なお、音声区間の判定結果として、音声区間の始端や終端が検出された時に、どのフレームが始端であるとか、どのフレームが終端であるとかの情報も記憶部２０に書き込むようにするものとする。

【００４８】このステップＳ６に示すような音声区間の判定を制御部１０が行う場合、連続して有音声部と判定されるような図５の（A）に示す音声信号に対しては問題ないが、図５の（B）に示すような音声区間Ｖの前部において無音声部Ｕに分離された短い有音声部ｖが存在する場合や、図５の（C）に示すように音声区間Ｖ終了後に無音声部Ｕが一旦入力され、続いて短い有音声部ｖが入力された場合などに、これらを音声区間として含めるかどうかという問題に対処できない。このような前後に付加された短い有音声部ｖは、音声区間として含めた方が良いものもあれば、誤って有音声部と判定された非定常ノイズなど、音声区間に含めない方が良いものもある。

【００４９】そこで、ステップＳ６の音声区間判定時に、以下に述べるような処理を追加する。先ず、問題となるのは、mフレーム未満しか連続しない有音声部の後に無音声部が入力され、その後再び有音声部が入力された場合の処理である。そこで、このmフレーム未満の有音声部のフレーム数とその後に続く無音声部のフレーム数をカウントするようにし、mフレーム未満の有音声部の後に、無音声フレームが連続して入力された場合に、音声区間と見なすか見なさないかを図６の（A）のような関係に基づいて判定する。すなわち、もし、m_{pre}フレーム（例えば３フレーム）未満の有音声部ｖの後にn_{pre}フレーム以上の無音声部Ｕが入力された場合は、そのm_{pre}フレーム未満の有音声部ｖは音声区間から除去するものとする。ただし、$m_{pre}<m$とする。また、ｘフレーム（m_{pre}フレーム以上mフレーム未満）の有音声部の後に、

【００５０】

【数４】

$$n_{pre}+\frac{n-n_{pre}}{m-m_{pre}}x \quad \cdots(4)$$

10

【００５１】フレーム以上の無音声が連続して入力されれば、そのｘフレームの有音声部も音声区間から除去する。つまり、図６の（A）に示すように、横軸を有音声部の入力フレーム数、縦軸をその後に続く無音声部のフレーム数としたとき、斜線部の関係の場合は、その有音声部を音声区間に含めず、斜線部より下の関係の場合は、その有音声部との間にはさまった無音声部を音声区間に含めるようにする。

【００５２】以上のような処理を追加することにより、図５（B）に示されるような音声区間Ｖの前部の短い有音声部ｖの扱いとして、挿入された無音声部Ｕが短い時には該短い有音声部ｖを音声区間に含め、長い時には該短い有音声部ｖを音声区間に含めないようにすることが可能となる。

【００５３】同様に、図５の（C）のような音声区間Ｖの最後に無音声部Ｕで分離された短い有音声部ｖが存在する場合の処理も同じように行う。すなわち、図６の（B）に示されるように、mフレーム以上の有音声部の後に発生した無音声部のフレーム数（縦軸）とその後に表れるmフレーム未満の有音声部のフレーム数（横軸）をカウントするようにし、無音声フレームが連続して入力された後に、mフレーム未満の有音声部ｖが入力された場合に、音声区間と見なすか見なさないかを判定している。つまり、図６の（B）に示す斜線部のような関係にあれば、その最後の有音声部分は音声区間から除去し、斜線部の下のような関係にあればその最後の有音声部分との間に挟まった無音声部を音声区間に含めるようにする。すなわち、n_{post}フレーム以上の無音声部Ｕが入力された後に、m_{post}フレーム（例えば３フレーム）未満の有音声部ｖが入力された場合は、そのm_{post}フレーム未満の有音声部ｖは音声区間から除去する。ただし、$m_{post}<m$とする。また、ｘフレーム（m_{post}フレーム以上mフレーム未満）の有音声部の前に、上記（４）式と同様の式、すなわち上記（４）式のn_{pre}をn_{post}、m_{pre}をm_{post}と変更した式によって表されるフレーム以上の無音声部が連続して入力されていれば、そのｘフレームの有音声部も音声区間から除去する。

【００５４】次に、制御部１０は、ステップＳ７に示すように、ステップＳ６で音声区間と判定されたフレームの前後に、さらに音声区間としてのマージンを付加し、実際の音声区間より少し長めの音声区間を抽出する。これは、ステップＳ６の検出誤り、すなわち音声区間の始端が遅れて検出されてしまったり、終端が速く検出されてしまうのを防ぐために行う。

【００５５】例えば、図４の（D）に示すように、音声区間の始端が検出された時点、すなわち、時刻ｔにフレームF_tが入力され、そのm時刻前のフレームF_{t-m}が音声区間の始まりとして検出されたときに、記憶部２０にフレームF_{t-m}が音声区間であると書き込むと同時に、そこからさらにｐフレーム前までのフレームに対しても

音声区間であったことを追加して書き込むようにすればよい。

【００５６】同様に、図４の（Ｅ）のような音声区間の終了が検出され、その終了フレームの判定結果が記憶部２０に書き込まれてから、さらにｑフレーム後まで音声区間とみなして、記憶部２０に書き込み続けるようにする。以上のようにして、ステップＳ６において、判定された音声区間のフレームの前にｐフレーム、後ろにｑフレームのマージンを付加したものが最終的な音声区間として記憶部２０に記憶されていくことになる。その際、音声区間の始端や終端の情報も、対応したものに変更しておくようにする。ここで、マージンｐ、ｑは一定値を用いたり、あるいは後述するような、環境雑音の平均エネルギーに応じて変動させた値を用いてもよい。

【００５７】次に、制御部１０は、ステップＳ８に示すように、音声区間の判定のためのパラメータを環境の変化に応じて更新する。音声区間の判定のためのパラメータ、すなわち環境雑音の推定パワースペクトルＮ（ω）や環境雑音の平均エネルギーＥｎ、さらに、有音声・無音声の判定のためのしきい値ｒや音声区間のマージンｐ，ｑは、環境の変化に応じて変動させることが、耐雑音性能の向上のために必要となってくる。そこで、これらのパラメータの更新は毎フレーム行うことにし、一つのフレームの処理が終了した時点で、次フレームの処理のために新しく更新を行うようにする。

【００５８】まず、環境雑音の推定パワースペクトルＮ（ω）と環境雑音の平均エネルギーＥｎの更新について説明する。これは基本的には、音声区間外において、環境雑音の推定パワースペクトルＮ（ω）や環境雑音の平均エネルギーＥｎを平均化したものを求めるようにする。その求めかたとしては、先ず、入力フレームから求められるパワースペクトルＳ（ω）からエネルギーＥｓを、

【００５９】

【数５】

$$Es = \sum_{\omega} S(\omega) \quad \cdots(5)$$

【００６０】のように求め、そして、前フレームにおいて求められた上記環境雑音の推定パワースペクトルＮ（ω）と上記環境雑音の平均エネルギーＥｎを用いて、例えば、

【００６１】

【数６】

$$N(\omega) := \frac{(h-1)N(\omega) + S(\omega)}{h} \quad \cdots(6)$$

【００６２】

【数７】

$$En := \frac{(h-1)En + Es}{h} \quad \cdots(7)$$

【００６３】のように更新する。

【００６４】ここで、ｈ（例えば、２０）は重み係数であり、更新前の値に（ｈ－１）／ｈ、入力フレームの値に１／ｈの重みをかけて加え合わせることによって、時間的に新しいフレームに重みをおいて平均化を行った結果が得られることになる。

【００６５】なお、上記の更新は、音声区間外のフレームにおいてのみ行うものとし、例えば、図４の（Ｂ）や、図４の（Ｇ）のように、ｎフレーム前から連続して無音声と判定されつづけ、しかも前フレームで求められた環境雑音の平均エネルギーＥｎに比べて入力フレームのエネルギーＥｓが急激に大きくなり過ぎないような入力フレーム、例えば、ＥｎとＥｓを比較したとき、

Ｅｎ／Ｅｓ＞０．５　　・・・（８）

を満たすような入力フレームにおいてのみ更新を行うようにする。つまり、有音声と判定されるフレームや、既に求めてある環境雑音の平均エネルギーから大きく離れたエネルギーをもつようなフレームにおいては、環境雑音のパラメータ更新を行わないようにすることで、ある程度安定した環境雑音の推定パワースペクトルＮ（ω）や環境雑音の平均エネルギーＥｎが得られることになる。

【００６６】このようにして得られた環境雑音の平均エネルギーＥｎに適応させて、他のパラメータを更新する方法に関して説明する。あるパラメータＰを、環境雑音の平均エネルギーＥｎが大きなときには大きく、Ｅｎが小さいときには小さくしたい場合、簡単には、パラメータＰを環境雑音の平均エネルギーＥｎやその対数に比例させて変動させる方法が考えられる。この際、パラメータＰの変動の範囲を制限するため、パラメータＰの上限下限を設定する。これを示したのが図７の（Ａ）である。横軸が環境雑音の平均対数エネルギーＨであり、この平均対数エネルギーＨは、

【００６７】

【数８】

$$H = 10\log_{10} En \quad [dB] \quad \cdots(9)$$

【００６８】のように求められる。また、縦軸がパラメータＰである。Ｐmax、Ｐminは、パラメータＰの下限上限を示している。また、比例定数は、例えば、パラメータＰを環境雑音がＨａ［ｄＢ］の場合にＰａ、Ｈｂ［ｄＢ］の場合にＡｂに設定したい場合、

【００６９】

【数９】

$$\frac{P_b - P_a}{H_b - H_a} \quad \cdots(10)$$

【００７０】のようにして求めることができる。このようにして、図７の（Ａ）のような関係を予め決めておけば、環境雑音の平均エネルギーＥｎからパラメータＰを求めることが可能となる。同様に、パラメータＰを、環境雑音の平均エネルギーＥｎが大きいときに小さく、環

境雑音の平均エネルギーE_nが小さいときに大きくしたい場合には、図７の（Ｂ）のような関係に基づいて求めれば良い。比例定数はやはり上記（１０）式で求められる。

【００７１】このような環境雑音の平均エネルギーE_nに適応させたパラメータＰの更新方法は、例えば、音声区間の前後に付加するマージンｐ、ｑのように、環境雑音が大きい場合には音声区間の判定の精度が悪くなるので大きくし、環境雑音が小さい場合には音声区間の判定精度が良いので小さくしたいというときなどに用いることができる。また、上記（１）式の重み係数αや、有音声・無音声の判定のためのしきい値ｒなどを、環境雑音が大きい時には大きく、小さいときには小さくしたい場合、あるいは逆に、環境雑音が大きいときには小さく、小さいときには大きくしたい場合などに適用することができる。

【００７２】以上、環境雑音の平均エネルギーE_nの更新方法と、該エネルギーE_nに応じた他のパラメータＰの更新方法に関して述べた。基本的には、あるパラメータＰを環境に適応させて変動させたい場合、パラメータＰと環境雑音の平均エネルギーE_nの対応関係である

$$P = f(E_n)$$

を予め決めておき、環境雑音の平均エネルギーE_nからパラメータＰを求めるようにすればよい。

【００７３】なお、パラメータＰを一定にすることは、$P_{max} = P_{min} = constant$とすることに対応する。

【００７４】次に、制御部１０は、ステップＳ８で示したパラメータ更新の例外として、音声区間が一定以上長く続くような場合に、ステップＳ９に示すように強制的なパラメータ更新を行う。

【００７５】上記ステップＳ８のパラメータ更新において、環境雑音の推定パワースペクトル$N(\omega)$や環境雑音の平均エネルギーE_nは、音声区間外のフレームにおいてのみ更新を行うものとしたが、例外として、音声区間がある一定以上長く続くような場合には、強制的に更新を行うようにしておく。つまり、図４の（Ｅ）のような状態が長く続き、音声区間として判定されるフレームがＱ（例えば５００）フレーム以上続いた場合に、上記$N(\omega)$や上記E_nを強制的に更新するようにする。これは、環境雑音が急激に大きくなった場合などに、誤って有音声と判定し続け、しかも、環境雑音の推定パワースペクトル$N(\omega)$や環境雑音の平均エネルギーE_nがこれに追従できないという問題が生じるのを避けるためである。このような処理を付加しておくことで、環境雑音の変動に対してさらに強くすることが可能となる。

【００７６】次に、制御部１０は、ステップＳ１０に示すように、最終的に音声区間を検出し、送信を行う。

【００７７】上述したステップＳ９までの処理が終了したとき、図２の記憶部２０には入力フレームF_tからｋフレーム前までのフレーム信号と、その時点までの判定

結果、すなわち、F_{t-m}、F_{t-m-1}、・・・F_{t-k}に対して付けられた音声区間の判定結果、及び音声区間の始端・終端情報が記憶されていることになる。ここで、ｋは音声区間として判定するために最低必要なフレーム数ｍに、音声区間の始端を前にずらすためのマージンｐの最大値を加えたものより大きく、かつ無音声の区間を判定するために最低必要なフレーム数ｎよりも大きいものとする。これは、音声区間の判定が行われるのが入力されてからｍフレーム後であり、しかもマージンが付加される場合は、さらにｐフレーム後となるため、そのフレームの判定が確定するのは入力されてからｍにｐの最大値を加えたフレーム分だけ遅れてからとなること、及び、音声区間の終了が検知できるのは入力されてからｎフレーム遅れてからとなることから要求される。

【００７８】そして、制御部１０は、この記憶部２０から、ステップＳ１０に示すように、必要な音声区間を検出して出力する。図１の区間指定スイッチ４が押された場合、図２の制御部１０に対してスイッチ信号が送られて来る。これには、ほとんど遅延がない。この区間指定スイッチ４のスイッチ信号の送られてくる様子を示したのが、図８の（Ａ）である。横軸が時間、縦軸がスイッチ信号であり、区間指定スイッチ４が押されている間は、“１”、押されていないときは“０”の信号が送られて来るものとする。これに対して、入力フレーム信号は図８の（Ｂ）に示すようにΔ時間毎に送られてくる。つまり、区間指定スイッチ４が押され出した時点t_sと区間指定スイッチ４が離された時点t_eは、図８の（Ｃ）に示されるように、あるフレームが入力されてから次のフレームが入力されるまでのΔの間に検知されるはずである。

【００７９】そこで、区間指定スイッチ４の押され出した時点t_sの直後に入力されたフレーム（図８の（Ｃ）の場合F_{t+1}）を区間指定スイッチ４のオンの起点とし、区間指定スイッチ４が離された時点t_eの直後のフレーム（図８の（Ｃ）の場合F_{t+4}）を区間指定スイッチ４のオフの起点とする。そして、図９の（Ａ）に示すように、区間指定スイッチ４のオンの起点のｋフレーム前から、区間指定スイッチ４のオフの起点のＬフレーム後ろまでの区間をスイッチ指定区間とする。このとき、各フレーム信号が入力される度に、区間指定スイッチ４のオン、オフの起点となるかどうかを調べることができ、これを基にして記憶されているｋフレーム前から入力フレームまでの信号がスイッチ指定区間に含まれるか否かを判定できるので、その判定結果も記憶部２０に書き込むことにする。この際、スイッチ指定区間の始端と終端の情報も記憶するようにしておく。そして、このスイッチ指定区間の中に、音声区間の始端が検出されれば、その音声区間の始端から終端までを必要な音声区間として図１０の（Ａ）に示す範囲で検出し、これを送信する。ここで、音声区間がスイッチ指定区間終了後も続

くような場合でも、図１０の（Ｂ）に示すように、その音声区間の終了までは送信するものとする。

【００８０】送信の仕方は、図１１に示すようにする。すなわち、送信すべき音声区間の信号を入力させてから、ｋ時刻遅れて送信する。時刻ｔにおいては、図２の記憶部２０に記憶されているｋ時刻前のフレームＦ$_{t-k}$を常に着目し、そのフレームが送信すべき音声区間のフレームであるときのみ送信し、そうでないときは送信しない。こうすることによって、送信すべき音声区間のフレームは、入力フレームＦ$_t$に同期して、ｋ時刻遅れで送信されることになるわけである。

【００８１】次に、送信すべきか否かの判断について述べる。時刻ｔにおいて着目するのは、ｋ時刻前のフレームＦ$_{t-k}$である。この時点において、記憶部２０には、フレーム信号Ｆ$_{t-k}$が記憶されていると同時に、これがスイッチ指定区間であるかどうか、音声区間であるかどうかも既に判定され記憶されていることになる。また、音声区間であれば、音声区間の始端・終端であるかどうかも記憶されていることになる。そこで、図１２に示すような状態遷移図にしたがって、着目フレームＦ$_{t-k}$がどの状態からどの状態に遷移するかを調べ、その状態遷移に応じて送信すべきかどうかを決定するようにする。

【００８２】先ず最初、状態(i)から始める。そして、区間指定スイッチ４のオンの起点が検知されるまで、状態(i)で自己遷移する。もし、区間指定スイッチ４のオンの起点が検知されれば、着目フレームＦ$_{t-k}$が音声区間の始端のフレームかどうかを調べ、ＮＯならば状態(ii)に、ＹＥＳならば状態(iii)に遷移する。状態(ii)に遷移した後は、着目フレームＦ$_{t-k}$が音声区間の始端のフレームかどうかを調べ、ＹＥＳならば状態(iii)へ遷移する。着目フレームが音声区間の始端でない場合は、さらに着目フレームがスイッチ指定区間の終了フレームかどうかを調べ、ＮＯならば状態(ii)で自己遷移し、ＹＥＳならば再び状態(i)に戻る。状態(iii)においては、着目フレームが音声区間の終端フレームかどうかを調べ、ＮＯならば状態(iii)で自己遷移する。もし、着目フレームが音声区間の終端フレームとなった場合は、その時点からｋ時刻前まで、つまりフレームＦ$_{t-k-1}$からＦ$_t$が入力された間に区間指定スイッチ４のオンの起点が検知されたかどうかを調べ、ＹＥＳならば状態(ii)に、ＮＯならば状態(i)に遷移する。そして、このような状態遷移において、状態(iii)への遷移時と状態(iii)からの遷移時、すなわち、状態(i)から状態(iii)、状態(ii)から状態(iii)、状態(iii)における自己遷移、状態(iii)から状態(i)、状態(iii)から状態(ii)の遷移時において、着目フレームＦ$_{t-k}$を送信するようにする。逆に、それ以外の遷移時には、送信を行わないようにする。

【００８３】以上のような処理により、着目フレームがスイッチ指定区間に入った後、音声区間の始端が検知されてから終端が検知されるまで送信が行われるようになる。もし、スイッチ指定区間が終了するまでに、音声区間の始端が検知されないような場合は、何も送信されないままとなる。（状態(ii)から状態(i)）。また、区間指定スイッチ４のオンの起点が検出されない限り、状態(i)から他の状態への遷移は起こらないため、一つのスイッチ指定区間に送信される音声区間は一つまでとなる。つまり、区間指定スイッチを一度押しただけで、２つ以上の発話を行ったとしても、送信されるのは最初に検出された音声区間だけとなる。ただし、発話終了後、すぐに、区間指定スイッチ４を押して、再び発話した場合に、最初の発話に対応する音声区間も二つ目の発話に対応する音声区間も送信するようにするため、状態(iii)から状態(ii)への遷移を設けておく。つまり、音声区間の信号は、入力時刻よりｋ時刻遅れてから送信されることを考慮して、音声区間の終端のフレームが入力されてから送信されるまでの間に区間指定スイッチ４のオンの起点が検出された場合は、状態(iii)から状態(ii)へ遷移するようにしておく。

【００８４】なお、着目フレームがスイッチ指定区間として終了していないにも関わらず、再びスイッチが押されるようなことがあれば、スイッチ指定区間を延長し、再度押された区間指定スイッチ４のオフの起点からさらにＬフレーム後ろまでを図９の（Ｂ）に示すように、新たなスイッチ指定区間とする。

【００８５】なお、上述した実施例となる音声区間検出装置は、音声通信システムに組み込むこともでくる。この場合、音声信号の送受信を行う制御部を図１の音声認識部５の代わりに設ければよい。このため、入力された音声信号の中から必要な音声区間だけを抽出し、送信することが可能となる。

【００８６】

【発明の効果】本発明に係る音声区間検出装置は、音声信号が入力され、検出範囲を指定するスイッチ操作に応じて音声区間を抽出して出力する音声区間検出装置において、上記入力音声信号を一定時間分記憶する記憶手段と、上記スイッチ操作で指定される検出範囲よりも広い範囲で上記記憶手段に記憶されている上記入力音声信号から一つだけ音声区間を抽出し、出力する制御手段とを有するので、有音声と無音声の判別による音声区間判定を常に行いながら、その判定結果と一定時間前までの入力信号を記憶しておくことで、スイッチが押されるより早いタイミングで発声が行われた場合でも、語頭を欠くことなく音声区間を検出することが可能と成る。又、一つのスイッチ指定区間に対して一つの音声区間だけを検出し送信するようにしたことで、話者が発声の少し前や、少し後に関係のない言葉を発声したとしても、これを音声区間として検出することがなくなるため、例えば音声認識装置に適用した場合などに誤動作を起こしにくくなる。

17

【００８７】さらに、入力信号を周波数分析して得られるパワースペクトルを用いること、及び、環境雑音の平均エネルギーに適応させてパラメータを更新することなどを行うことで、環境雑音下においても安定した音声区間検出が行えるようにすると共に、その適応化を音声区間の判定と平行して常に行っておくようにすることで、スイッチを押してから雑音環境の観測を行う必要が特になくなるため、ユーザが発声するのを待たされることもなくなる。つまり、ユーザは、単純にスイッチを押せば一回だけ発話することができるということだけを覚えておけばよく、その発話とスイッチのタイミングをそれほど意識しなくてよいものとなる。しかも、システムは、必要な音声区間だけを検出することが可能となる。

【００８８】そして、この音声区間検出装置を音声認識システムや音声通信システムに組み込めば、入力信号のなかから認識処理を行うべき音声区間や送信すべき音声区間を抽出することができるようになる。

【図面の簡単な説明】

【図１】本発明の音声区間検出装置を音声認識システムに適用した例を示したブロック図である。

【図２】本発明の実施例の音声区間検出装置のブロック図である。

【図３】上記実施例の音声区間検出装置の動作を説明するためのフローチャートである。

18

【図４】上記実施例の音声区間検出装置の制御部の音声区間の判定について説明するための図である。

【図５】無音声部によって音声信号の一部が分離された状態を示す図である。

【図６】図５に示した分離された音声区間の判定基準を説明するための図である。

【図７】パラメータ更新における環境雑音のエネルギーと環境雑音のパラメータの関係を示した図である。

【図８】スイッチ信号と入力フレーム信号の時間的関係を示した図である。

【図９】スイッチ指定区間を説明するための図である。

【図１０】スイッチ指定区間と音声区間の関係を示した図である。

【図１１】音声区間信号の送信の仕方を説明するための図である。

【図１２】音声区間信号の送信の判定について示した図である。

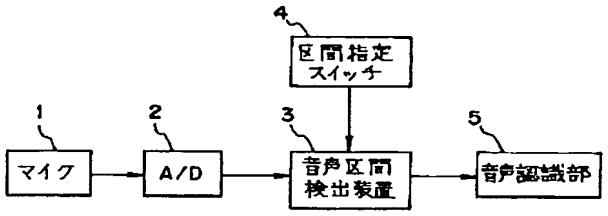
【符号の説明】

３　音声区間検出装置

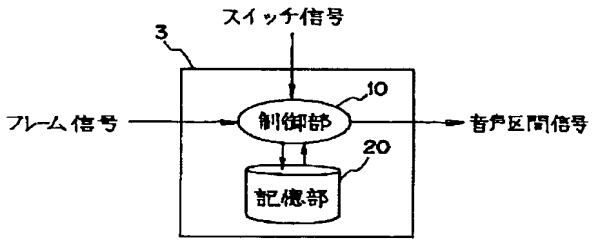
１０　制御部

２０　記憶部

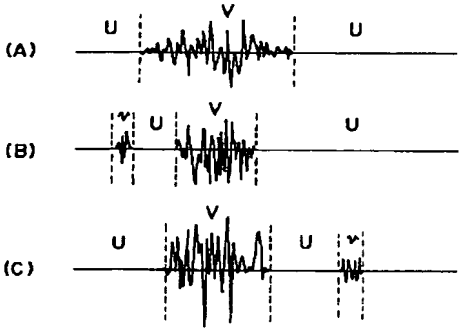
【図１】



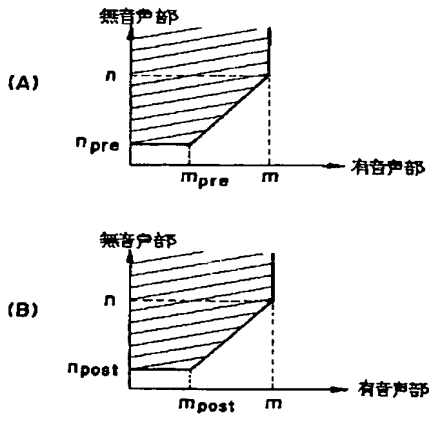
【図２】



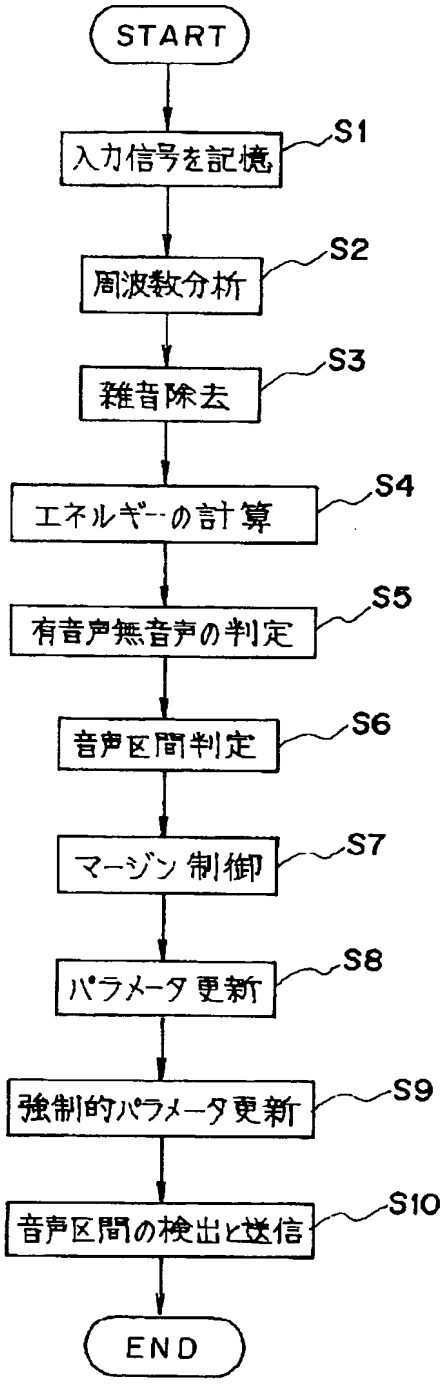
【図５】



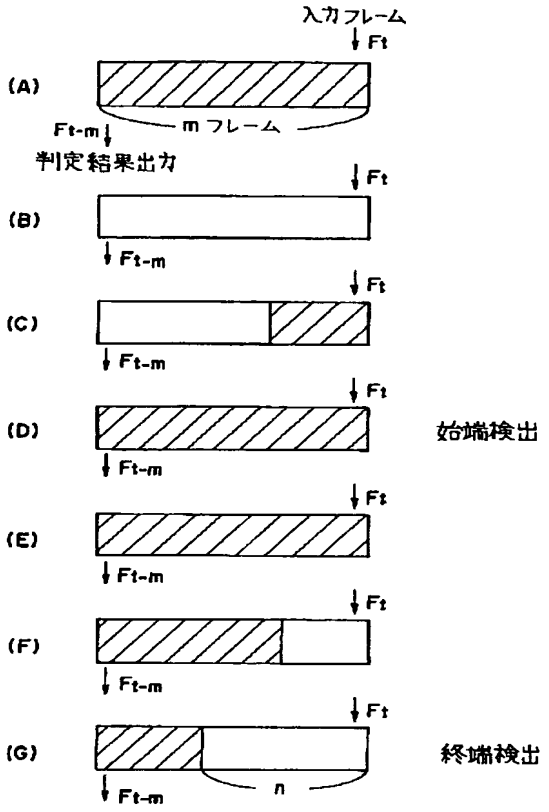
【図６】



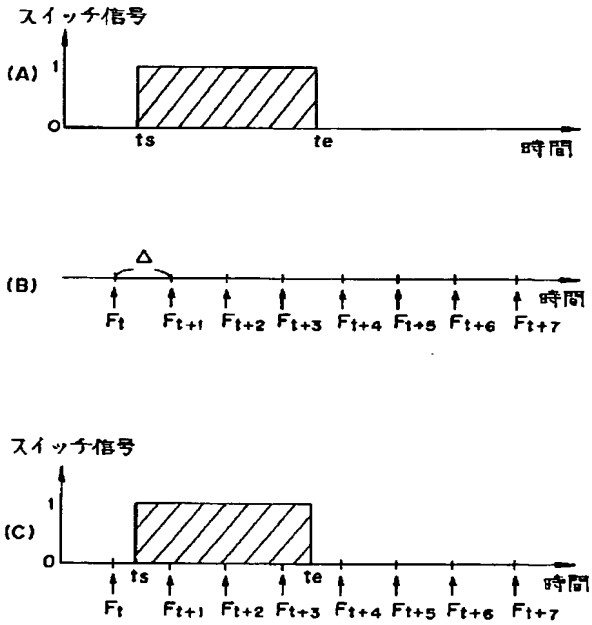
【図3】


START
入力信号を記憶
S1
周波数分析
S2
雑音除去
S3
エネルギーの計算
S4
有音声無音声の判定
S5
音声区間判定
S6
マージン制御
S7
パラメータ更新
S8
強制的パラメータ更新
S9
音声区間の検出と送信
S10
END

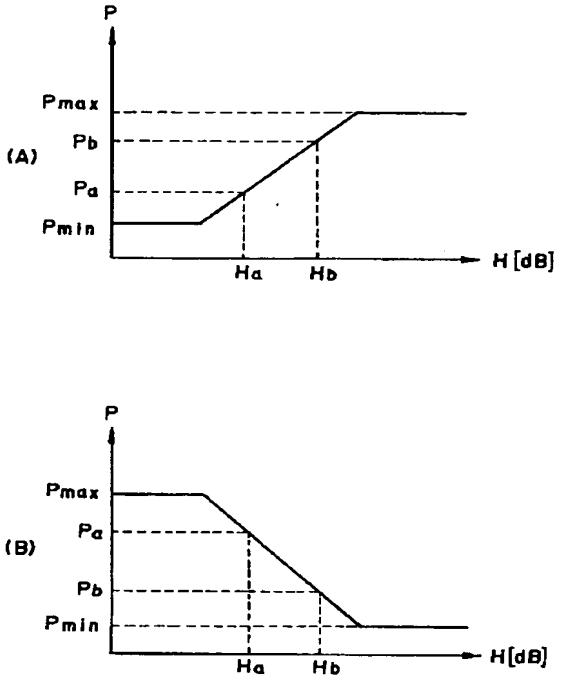
【図4】


入力フレーム
Ft
(A)
Ft-m
mフレーム
判定結果出力
(B)
(C)
(D)
始端検出
(E)
(F)
(G)
終端検出
n

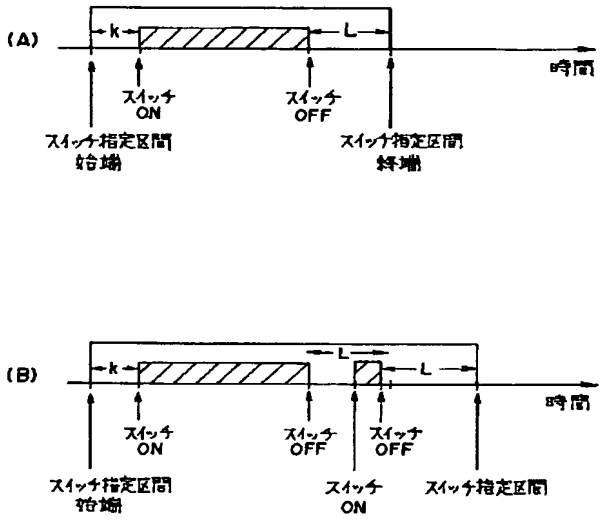
【図8】


スイッチ信号
(A)
1
0
ts
te
時間
(B)
Δ
Ft
Ft+1
Ft+2
Ft+3
Ft+4
Ft+5
Ft+6
Ft+7
時間
スイッチ信号
(C)
1
0
ts
te
時間

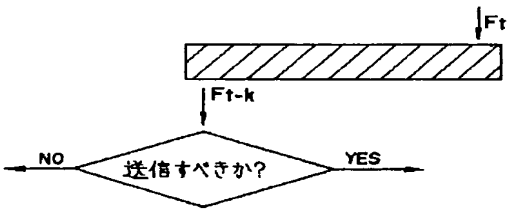
【図7】


P
Pmax
Pb
(A)
Pa
Pmin
H[dB]
Ha
Hb
P
Pmax
Pa
(B)
Pb
Pmin
H[dB]
Ha
Hb

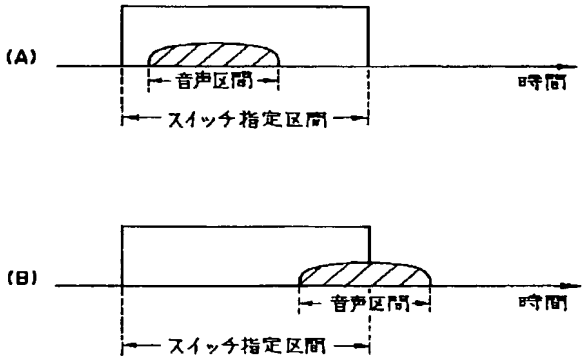
【図9】


(A)
k
L
時間
スイッチON
スイッチOFF
スイッチ指定区間始端
スイッチ指定区間終端
(B)
k
L
L
時間
スイッチON
スイッチOFF
スイッチOFF
スイッチ指定区間始端
スイッチON
スイッチ指定区間

【図11】


Ft
Ft-k
NO
送信すべきか?
YES

【図10】


(A)
音声区間
時間
スイッチ指定区間
(B)
音声区間
時間
スイッチ指定区間

【図１２】